Determining Syntactic Features of English Registers with Linear Support Vector Machine

A.Z. Vasquez, R. Porter
{avasquez1, reeseporter} at ufl dot edu

UF | UNIVERSITY of FLORIDA

Introduction

The process of training a machine can be lengthy and is a recurring task in computational linguistics and in the larger domain of data science. Because of this fact, research in algorithm optimizations is evermore relevant. For machines studying English registers, it would be optimal to avoid spending processing time on learning predictors that prove to be ineffective. By this study, we explore a set of predictors of English register and propose a subset of those predictors that is most useful for this task.

Which syntactic features are most useful in computationally identifying a text's register (conversational, academic, news) in standard English?

The following features are claimed to be distinctive between the English registers in question:

By *Wang, et. al*
- dependency distance

By *Biber, et. al*
- adverbial clause frequency
- modal/semi-modal frequency
- perfect aspect frequency
- progressive aspect frequency
- personal pronoun frequency

Datasets

10,000 samples of each English register were collected for a total of 30,000.

Conversational Register

Convokit's movie-corpus was queried for 10,000 random conversations.

News Register

The NLTK Reuters corpus has news articles. 10,000 news articles were queried randomly.

Academic Register

For samples of academic register, a random keyword was retrieved from the Brown Corpus which became the search query on arXiv.org, a corpus of academic articles. The summaries of the first 100 hits were aggregated. This process continues until 10,000 article summaries are gathered.

Methods

1. Gather 10,000 samples of each register. (See Datasets below)
2. Tag each sample with each selected feature.
3. For each combination of features, train a Linear Support Vector Machine model on a training set to classify the registers. Use a train-test split of 50:50.
4. Test and evaluate models on their F1-score and precision.
5. Compare models and record observations.

References

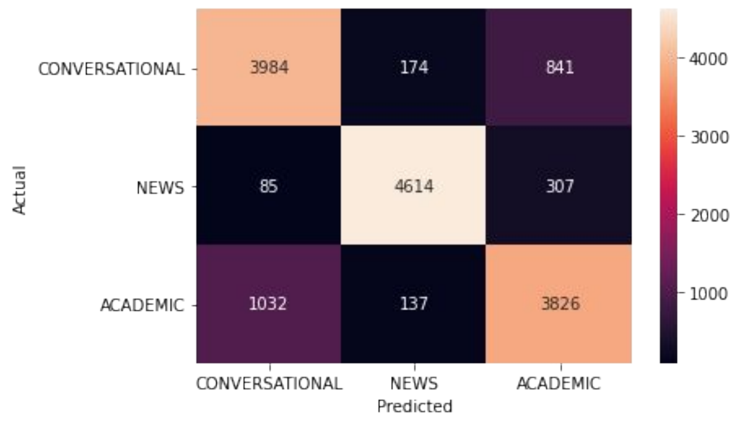
Biber, D., Conrad, S., & Leech, G. N. (2002). *Longman student grammar of spoken and written English*.

Wang, Y & Liu, H. (2016) *The effects of genre on dependency distance and dependency*

Results and Observations

Of the 63 classifiers trained, the **greatest F1-score (~0.829)** belonged to the model trained on **dependency distance, adverbial clause frequency, perfect aspect frequency, progressive aspect frequency, and personal pronoun frequency.** Shown below is the model's confusion matrix with counts for however many times the classifier classified each document as each register.

Interestingly, Academic and Conversational documents get confused often, as shown by the purple squares at the intersections of Academic and Conversational registers.We intuitively hypothesized that Conversational and Academic registers would be more often confused, since we associate those registers with factual, informative sentences in contrast with Conversational register.

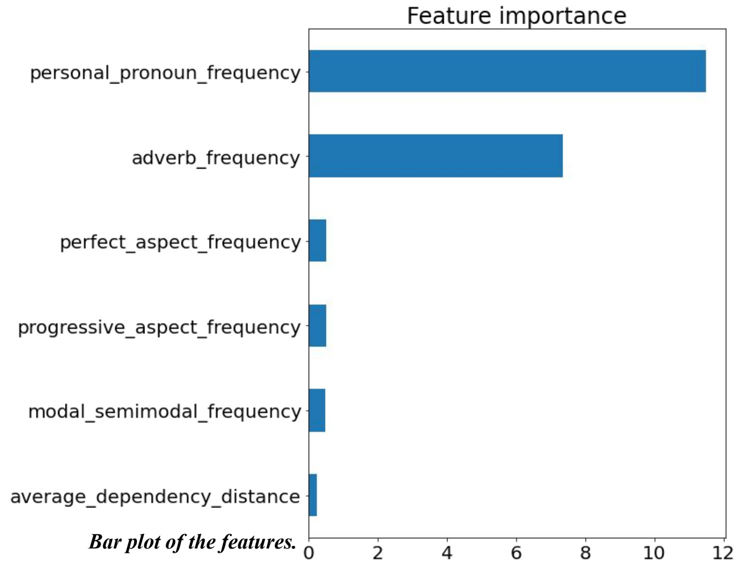


Confusion matrix of the best-performing model.

For the model trained on all the features, the coefficients correlating these features and each register is shown in the table below. Coefficients highlighted in green **correlative positively** with that register while coefficients highlighted in blue **correlate negatively** with that register. News registers are characterized by many adverbial clauses and a lack of personal pronouns. Conversational registers are characterized by many adverbial clauses and personal pronouns. Academic registers are characterized by their lack of both adverbial clauses and personal pronouns.

Register	Average Dependency Distance Coefficient	Adverbial Clause Frequency Coefficient	Modal/Semi-modal Frequency Coefficient	Perfect Aspect Frequency Coefficient	Progressive Aspect Frequency Coefficient	Personal Pronoun Frequency Coefficient
News	0.271	2.959	-0.489	-0.683	0.480	-3.041
Conversational	-0.407	6.612	0.902	-0.041	0.382	14.944
Academic	-0.109	-12.538	0.057	0.859	-0.719	-16.500

The models with the lowest F1-scores were generally the ones trained only on a single feature. In *descending order of F1-score*, these features are: Personal pronoun frequency (F1 = 0.644), dependency distance (0.534), adverbial clause frequency (0.464), progressive aspect frequency (0.373), perfect aspect frequency (0.364), and modal/semi-modal frequency (0.343).



Bar plot of the features.

Discussion

From the set of features studied, the best subset for identifying between News, Conversational, and Academic registers of English is that with **dependency distance, adverbial clause frequency, perfect aspect frequency, progressive aspect frequency, and personal pronoun frequency.**

In addition, syntactic generalizations have been made about these three registers as depicted by the weights assigned to each feature by the classifier.

From this study, a classifier has learned to accurately differentiate between the three English registers in question. As a result, the defining syntactic characteristics of each English register has been identified and shown to be useful in this task.